Research Report
ETS RR–11-31

The Single Group With Nearly Equivalent Tests (SiGNET) Design for Equating Very Small Volume Multiple-Choice Tests

Mary C. Grant

July 2011

The Single Group With Nearly Equivalent Tests (SiGNET) Design for Equating Very Small Volume Multiple-Choice Tests

Mary C. Grant
ETS, Princeton, New Jersey

July 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Technical Review Editor: Jim Carlson

Technical Reviewers: Skip Livingston and Sooyeon Kim

Abstract

The *single group with nearly equivalent tests* (SiGNET) design proposed here was developed to address the problem of equating scores on multiple-choice test forms with very small single-administration samples. In this design, the majority of items in each new test form consist of items from the previous form, and the new items that were administered as unscored items in the previous form. Each form is equated using data from examinees who took the previous form. As the equating is a single-group design, with the 2 forms having a large number of overlapping items, the size of the equating sample can be much smaller than in other designs.

Key words: equating, single-group design, small sample equating, testlets

Table of Contents

List of Figures

In the equating of test forms, if the samples used to derive the equating relationship between the new and reference forms are large and representative of the population, then the equating would likely have less sampling error and bias. Moreover, several sources on equating admonish practitioners about the problems of equating with small samples. For example, Kolen and Brennan (2004) recommend a rule of thumb of 400 examinees per form for linear equating and 1,500 examinees per form for equipercentile equatings when using either a nonequivalent groups with anchor test (NEAT) design or random equivalent groups. They go on to state that these numbers can be modified depending on a number of factors, including the shapes of the distributions, the degree of equating precision required, and the effects of smoothing for equipercentile equatings. When Kolen (1985) examined samples of 100 and 250, he found standard errors of equating, derived without the normality assumption, to be insufficiently accurate with a sample size of 100. On the other hand, Skaggs (2005), who studied equating using samples ranging from 25 to 200 in an equivalent groups design, concluded that although no equating (the identity equating function) was probably preferable to equating when samples are as small as 25, for samples in the range of 50 to 75, equating was preferable to no equating.

The effectiveness of smoothing small-sample distributions was investigated by Livingston (1993), who examined the efficacy of log-linear presmoothing with samples of 25, 50, 100, and 200 examinees, using the NEAT design. He found that presmoothing significantly reduced equating error and that the reduction was greatest for the smallest samples. He also found that equating using presmoothing was about as effective as unsmoothed equating using samples twice as large. In another study, no equating and linear equating were compared to equipercentile equatings using unsmoothed, presmoothed, and postsmoothed distributions for five ACT assessment tests (Hanson, Zeng, & Colton, 1994). It was found that equipercentile equating with small samples was significantly improved when the sample distributions were smoothed, however, neither smoothing method, pre- or post-, stood out over the other.

Not everyone has agreed that smoothing is necessarily the answer for equating with small samples. Holland, Dorans, and Petersen (2007) and Petersen (2007) stated that although smoothing helps in equating for moderate sample sizes, it still might not be much help for very small samples, particularly when it is unclear to what extent the small sample is representative of the larger population. Kim, von Davier, and Haberman (2006) noted that if the samples are very small, then log-linear smoothing, a nonlinear process, may introduce a sampling bias that can

counteract any gain that might occur due to a reduction in the standard error of equating. They proposed a compromise between the identity function and an estimated-equating function based on a small sample, referred to as a *synthetic linking function*. This synthetic function was determined to be preferable for moderately small samples such as 50 and 100; however, the identity function seemed preferable for very small samples ($N \leq 25$). The authors also suggested that the synthetic linking function might be a good choice when the test specifications are well defined and the test forms are nearly parallel.

Nearly everyone who does equating has an idea of what constitutes a small sample. However, a small sample for one testing program may be an adequate sample for another program. There are those who may be concerned about having to equate with only 300 examinees in the new form sample. On the other hand, there are those who would be grateful to have as many as 300 examinees in an equating sample. In programs that consistently have small equating samples, the point where psychometricians start to be really concerned tends to be somewhere near 100 examinees per sample, whereas the research has indicated that samples of 50 to 75 might be minimally adequate for some designs. Hence, in this paper, a working definition of a small equating sample is fewer than 50 examinees. Thus, although textbooks and papers on equating warn of the problems inherent in equating with small samples, the question is: What can be done in ongoing testing programs to equate a new form when, administration after administration, the examinee volume is very small (≤ 25)?

The difficulty is usually with the new form sample, that is, the group taking the new form at its first administration. For the reference form sample, most programs are able to accumulate sample volume over two or more administrations to achieve an acceptable sample size. In general, the typical multiple-choice test form equating uses nonequivalent groups and an internal-anchor (or common-item) equating method, such as Tucker, Levine, chained linear, or chained equipercentile. If a form is administered more than once, the volumes from multiple administrations of the reference form can be combined for the equating, but very few programs are in a position to delay score reporting in order to accumulate examinee volume for the new form sample over two or more administrations.

The SiGNET Design

For an ongoing testing program that is in the position of having to administer tests to fewer than 50 examinees at each administration, the choices are limited: (a) discontinue the test,

(b) continue administering the same form over and over for a long period of time, or
(c) introduce a new form, regardless how imperfect or inaccurate any equating procedure might be. In a relatively high-stakes testing program, especially one with established passing scores, none of these options is good. There are circumstances where a program is faced with the dilemma of equating with these small samples or not having scores to report. To alleviate this dilemma, a design was developed so that new forms could be introduced periodically and adequately equated. This design was originally referred to as the *testlet design for equating with small volumes* or the *testlet equating design*. It has more recently been christened by Dorans (personal communication, 2008) as the *single group with nearly equivalent tests design* or *SiGNET design*, which avoids confusion with any other equating design that uses testlets. This equating design was initially created for a testing program that included a number of multiple-choice tests that were given several times per year with administration volumes of five to 50.

Most previous attempts to deal with the small-sample equating dilemma have been to treat the data after collection (e.g., smoothing distributions) or to alter the equating method. The approach taken in the SiGNET design is to change the data collection design of the equating rather than to alter the method(s) used by combining some of the stratagems that are known to strengthen small-sample equatings. One of these stratagems is administering both the reference and new form to the same group of examinees so that the equating is a single-group design; another is having a large overlap of items in the two forms, which makes the two forms nearly equivalent in difficulty as well as in content.

A single-group equating requires the same group to take both the reference form and new form of the test. If the forms are given separately in a single operational administration, problems can arise. Most examinees do not want to take two full tests when only one is required. The time necessary to administer two complete and separate forms can be prohibitive. Also, there are practice and fatigue effects that can interfere with the equating. However, if the stratagem of having a large overlap of items or a large number of common items between the two forms is applied to those same forms and administration, that is, the two forms are administered together as if they were a single test form without repeating those items that are in common, then (a) the examinees will be unaware that they are taking two forms of the test, (b) the time needed to administer both forms will not be substantially greater than the time required to administer only one form of the test, and (c) there would not be any practice or fatigue effects from repeated

testing. Thus, combining single-group equating with a large overlap of items between the reference form and the new form addresses the problem of administering two forms at once.

The SiGNET design is a single group design that depends on the items in the test form being arranged in testlets. Although the word *testlet* carries different and specific meanings in some other measurement contexts, in this context, the primary characteristic of the testlets is that each testlet, to the extent possible, is a mini version of the total test. In the SiGNET design, the part of the test form that contributes to the examinees' reported scores consists of four or five testlets. In addition, there is a testlet that does not contribute to the examinees' scores on that form. This additional testlet will be part of the scores on the next form. In Figure 1, Form A consists of six testlets. Testlets 1–5 are included in the computation of examinees' reported scores for Form A. Testlet 6 is unscored in Form A, that is, that testlet is not included in the computation of examinees' reported scores for Form A. In Form B, the testlets contributing to the examinees' reported scores are Testlets 2–6. Because Testlets 2–6 are included in Form A, Form B scores can be equated to Form A scores using those examinees who took Form A. Form A only needs to be administered enough times to accumulate a sufficient sample for a single group equating where there is a large overlap of items. Figure 1 shows the scheme through Form D. Continuing in the same way, by the time Form F is assembled, there will be a completely different set of scored items as compared to those in Form A.

Form A	Form B	Form C	Form D
Testlet 1	Testlet 7*	Testlet 7	Testlet 7
Testlet 2	Testlet 2	Testlet 8*	Testlet 8
Testlet 3	Testlet 3	Testlet 3	Testlet 9*
Testlet 4	Testlet 4	Testlet 4	Testlet 4
Testlet 5	Testlet 5	Testlet 5	Testlet 5
Testlet 6*	Testlet 6	Testlet 6	Testlet 6

*Not included in computing examinees' reported scores.

Figure 1. **Test forms arranged in testlets for four forms in the SiGNET design.**

In Form B, Testlet 6 replaces Testlet 1 in computing reported scores and Testlet 7 is included as the unscored testlet that will be part of the next form, Form C (see Figure 1). Although Testlet 7 actually replaces Testlet 1 in Form B, it is Testlet 6 that will replace Testlet 1

in the computation of examinees' reported scores for Form B. Therefore, it is Testlet 6, not Testlet 7, which should most closely match Testlet 1 in both difficulty and content. Stated generally, the principle is that the unscored testlet in any form should most closely match the testlet that it will replace in the computation of examinees' reported scores.

Figure 1 shows the testlets as blocks of items that are added and deleted from the test form. However, to the extent possible, the items from each testlet should be scattered throughout the test form. If the items in each testlet were administered in blocks, there could be some consequences for equating. First, if the test was at all speeded, the last block in the test would seem more difficult than it actually was, because of fewer responses on the last several items of the test. Secondly, with the testlets administered as blocks, it would be much easier for the examinees to distinguish the testlets, themselves, and possibly play guessing games as to which are the unscored items. Neither of these possibilities is desirable. However, if the items of each testlet were scattered throughout the test form, not only would the problems of blocked items be avoided, but there would be the added advantage of being able to arrange the items based on other considerations, such as item difficulty, item content, or item chronology. See Figure 2 for an example of testlet items scattered throughout the test.

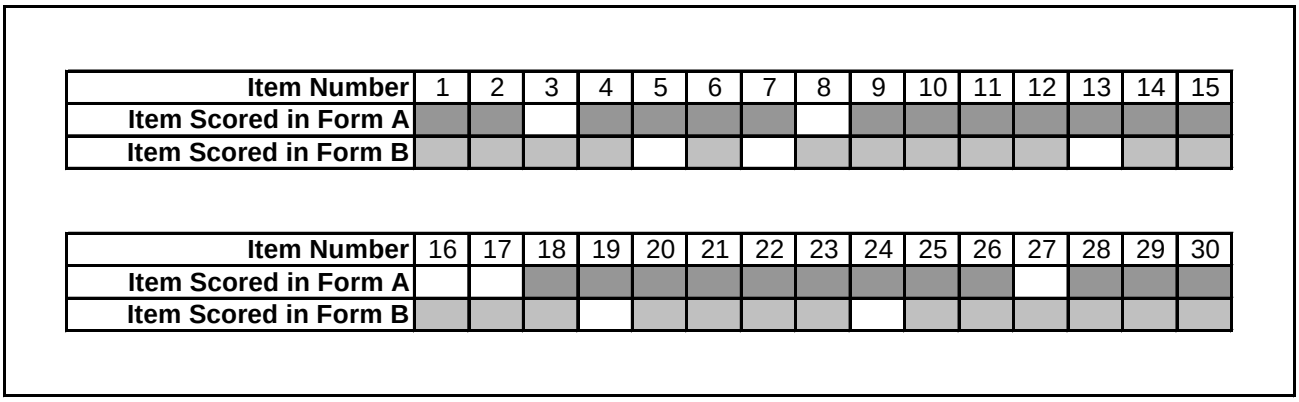

Figure 2. **Example of two forms administered as a single form with the unscored items in each form unshaded.**

Note. The unscored items in each form are scattered throughout the test.

In the substitution of items from Testlet 1 with the items from Testlet 7, the items from Testlet 7 should occupy the positions of the items that were deleted as part of Testlet 1. Although there might have to be some slight adjustment in item position to accommodate tests that have

sets of unequal length, this avoids a lot of reshuffling of items and allows the items in Testlets 2–6 to be in the same positions as they were in Form A, another advantage for equating.

SiGNET for Test Assemblers

Until this point, this design has been described primarily from the viewpoint of those who have to equate the forms. However, from the perspective of those who develop or assemble the forms, the design looks more like what is pictured in Figure 3. The first form consists of an operationally scored set of items consisting of four or five testlets and one unscored set of items consisting of a single testlet.

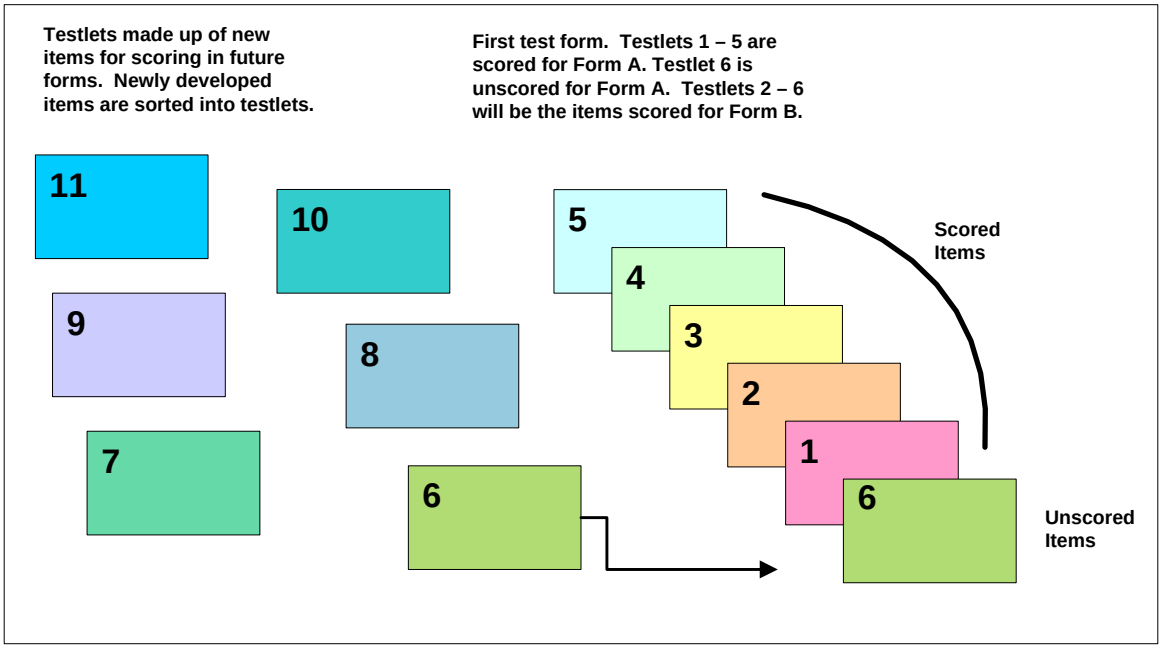

Figure 3. **Creating the first form comprised of testlets for the SiGNET design.**

Although the diagrams in most of the figures show the testlets as discrete entities, with the items for each testlet in a block, including the unscored testlet, it is better if the items for each testlet are actually scattered throughout the test. This arrangement allows for items to be in the same position in each test form in which the testlet is used. Figure 2 shows an example of the first thirty items with the items scored in each form shaded. The items that are unshaded for Form B will be replaced in the administered Form B with the new items of Testlet 7. This does entail more "bookkeeping" but it yields rewards in security (not being able to identify the

unscored block) and for equating because neither the reference form nor new form will have a block of items at either the beginning or end of the test.

From then on, the transition from form to form looks like Figure 1, to the person assembling the test. For each new form, the new item testlet replaces the testlet being deleted, including the item positions in the test. However the new testlet matches most closely in content and difficulty the remaining testlet next in line to be replaced. So, in Form A, Testlet 6 is the unscored testlet and it most resembles Testlet 1. Then, in Form B, Testlet 7 replaces Testlet 1, but most closely resembles Testlet 2. And, in Form C, Testlet 8 replaces Testlet 2, but most closely resembles Testlet 3.

A form is administered for as many administrations as necessary to accumulate sufficient volume for a single-group equating of two forms that have most of their items in common. Depending on the number of examinees at each administration, that length of time to accumulate sufficient volume will differ. If there are approximately 15 examinees at each admin, then equating can take place after approximately five (or more) administrations. If one assumes that a minimum of 75 examinee records would be required for an equating, then for any test with over 25 examinees per administration, a “new” form could be introduced about every fourth administration. For tests with fewer than 15 examinees per administration, a new form might be introduced every tenth administration. By the time one reaches the situation where there are fewer than ten examinees per administration, a “new” form could be introduced only about once every two to three years, assuming that there were at least five administrations in a year. That is just about the limit for this design. Replacing one fourth to one fifth of the test every few years would just about keep most tests current, but not much more.

There are two approaches for creating Form A depending on circumstances: Creating Form A as the initial form of a new test or creating Form A for an already-existing test using a possibly already-administered form.

The simplest situation is when there is a new test and it is known that volumes will be small. Then, the first form of the test can be set up as Form A in the SiGNET design. In other words, Form A is built entirely from an item pool and the five operational testlets are constructed so as to be the final form and Testlet 6 is constructed to be the unscored testlet that will replace Testlet 1 in Form B. Creating the initial form as the first form of a new test starts by building the operational test form according to test specifications and then dividing the items into Testlets 1–

5; Testlet 6 is constructed to closely match Testlet 1 in content and difficulty. The remaining items in the item pool will be used to create additional testlets for use in future new forms.

The second way to build an initial form for the SiGNET design is from an existing form of a test. For example, suppose there is an existing test for which the per-administration volume drops below that "critical number" for a NEAT design. Another example might be that the program introduces a new test with an already-assembled first form, but the expected number of examinees doesn't show up and it doesn't look like they will, at least not for the foreseeable future.

That existing form can be turned into a SiGNET design form fairly easily. First, the final length of the operational test needs to be determined, usually less than the number of items in the existing form. Then the number of items per testlet can be calculated. The items in the existing test form are then sorted into those testlets. Any extra items are put into a *discard testlet*. Ideally, these extra items should be the least desirable items from the existing form. The items in the existing test minus the items in the discard testlet are the operational form, Form A (see Figure 4). A testlet comprised of new items that will not be scored should be added to finish Form A (see Figure 3). To obtain the operational scoring and conversions for the new Form A, all that has to be done is a single group equating using the scores of examinees who have taken the old existing test form and equate the scores on the operational Form A items to the scores on the existing test form. This equating can be conducted before the first administration of Form A. From this point on, there is no difference from the Form A that was assembled from an item pool as the initial form.

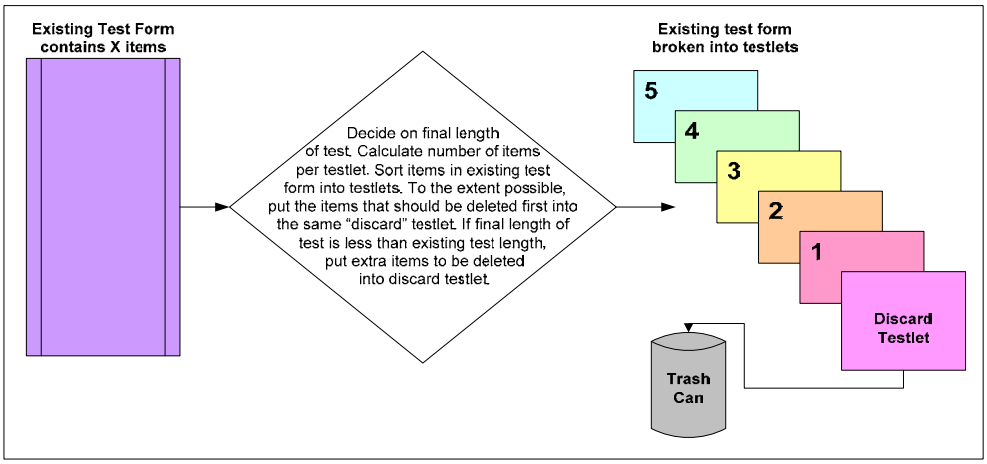

Figure 4. **The first steps in creating a form to be used with the SiGNET design from an existing test form.**

Discussion

In the one study that has compared the SiGNET design to NEAT design, Puhan, Moses, Grant, and McHale (2009) showed that especially for very small sample sizes, the SiGNET design was an improvement over the NEAT design. They also pointed out that for any program that requires preequating, such as some iBT applications, the SiGNET design is useful, because there is no wait to accumulate data for equating.

It should be noted that although the SiGNET design helps with equating and even with the evaluation of new items for subsequent forms, the evaluation of item performance at the first administration of the first form would be a problem. An item analysis run with 12 examinees doesn't yield much information. On the other hand, the design does not make things any more difficult than they were before.

The SiGNET design is extremely flexible. The testlets do not have to be unchanging collections of items. The testlets are primarily a means of keeping track of the items as each form rolls over to the next—Form A to Form B to Form C, etc. The only requirement is that if the item is to be scored in a form, for example Form B, it must have been administered to the examinees who were scored on the previous form, in this example, Form A.

Therefore, if an item is found to be flawed or becomes outdated, it does not have to remain in its original testlet and be administered but not scored through several forms until that testlet is cycled out. Because the equating is done using a single group and every examinee has taken all items in the reference form and in the unscored testlet, the new form can consist of any collection of items in that total set of items. In other words, all of the items in the testlet due to be dropped do not have to be dropped from the next form and all of the currently unscored items do not necessarily have to be part of the scored operational test in the next form. In addition, any item in any of the other testlets due to be kept in the next form could be dropped from the next form.

For example, suppose that while administering Form A, an item in Testlet 3 became outdated and had to be *not scored*. The testing program would not have to wait until Form D was introduced to get rid of the item. If there was an item in Testlet 1 that covered sufficiently similar content that it could be substituted for the outdated item, the operational items of Form B could consist of Testlets 2–6 minus the outdated item from Testlet 3 plus the item substitution from Testlet 1. Those items, as Form B, would be equated to the scored items in Form A (see Figure 5.)

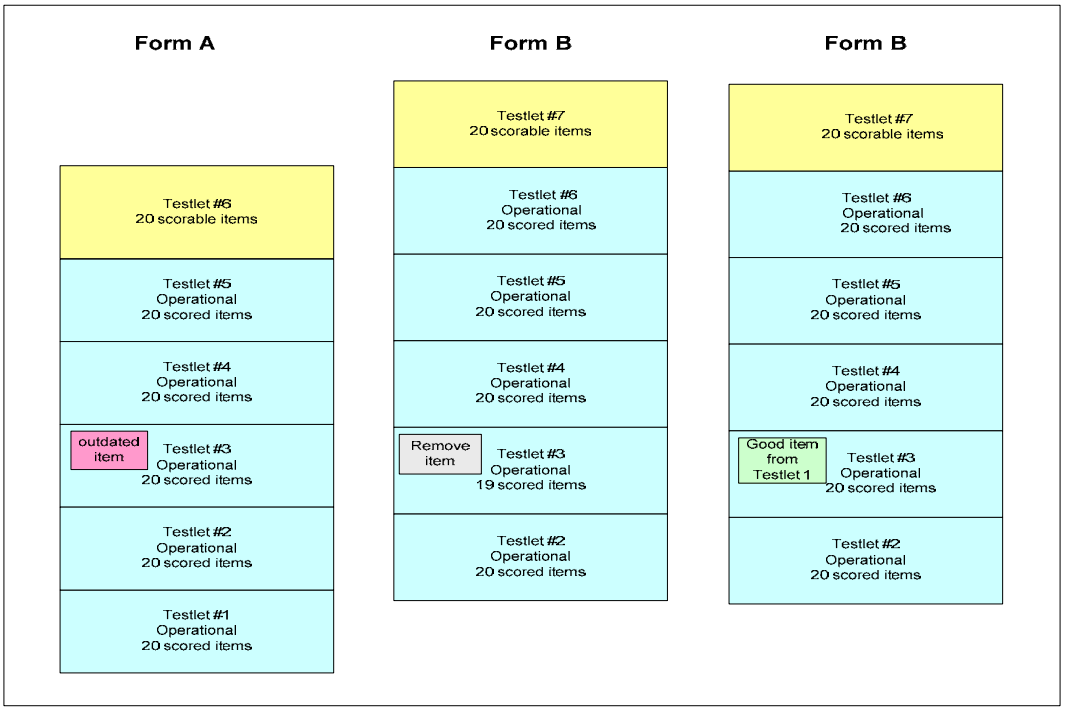

Figure 5. **A diagram of the process for replacing a flawed or outdated item.**

Note. The item substitution for Form B and the equating are completed using the Form A sample before Form B is assembled and administered**.**

If a flawed or outdated item cannot be replaced immediately by substitution, it can be carried into later forms of the test as a not-scored item—but it does not need to go through a full cycle of forms before it is replaced. Using the procedures outlined above, the item could be replaced in the next administered form with an additional item in the next new unscored testlet being prepared.

Conclusion

The SiGNET design offers a viable alternative for equating with very small administration volumes because the examinee records are accumulated until there are enough people in the sample for a reasonable single-group equating. Scaled scores on the scored can be reported to examinees shortly after each administration. The process can continue, replacing one testlet every time there are sufficient examinees for equating using a single group with equated scores ready each time a new form is introduced. Consequently, the SiGNET design will fit very nicely into an ongoing testing program with very small volume tests.

There are several strengths in the SiGNET design. The design allows for a single group equating, which means that a relatively small sample should result in an adequate equating. The format of the test form is such that examinees take both forms as a single test and cannot distinguish the two forms. The examinees cannot tell which items are being scored and which are not being scored. There is a 75–80% overlap of items in the two forms. Although the basis of the SiGNET design for small samples is that examinees take both the new and reference forms of the test in a single sitting, all of the examinees do not have to be tested at the same time. They may be tested over several operational administrations without delaying the reporting of scores.

Besides the facility to enable the equating of test form with very small samples, the SiGNET design has additional advantages:

- It is relatively easy to switch to the NEAT design when administration volumes increase to an acceptable size.
- There are some easy ways to replace flawed items that are not scored with new items without waiting for the testlets to run their full cycle.
- The design can function with a relatively small item pool and budget.

References

Hanson, B. A., Zeng, L., & Colton, D. (1994). *A comparison of presmoothing and postsmoothing in equipercentile equating* (ACT Technical Report No. 94-4). Iowa City, IA: ACT.

Holland, P. W., Dorans, N. J., & Petersen, N. S. (2007). Equating test scores. In C.R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Volume 26: Psychometrics* (pp. 169–201) Amsterdam, the Netherlands: Elsevier.

Kim, S., von Davier, A. A., & Haberman, S. (2006). *Equating with small samples* (ETS Research Report No. RR-06-27) Princeton, NJ: ETS.

Kolen, M. J. (1985). Standard errors of Tucker equating. *Applied Psychological Measurement, 9*, 209–223.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Livingston, S. A. (1993). Small-sample equating with log-linear smoothing. *Journal of Educational Measurement*, 30, 23–29.

Petersen, N. S. (2007). Equating: Best practices and challenges to best practices. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 59–71). New York, NY: Springer Science + Business Media.

Puhan, G., Moses, T. P., Grant, M. C., & McHale, F. (2009). Small-sample equating using a single-group nearly equivalent test (SiGNET) design. *Journal of Educational Measurement, 46,* 344–362.

Skaggs, G. (2005). Accuracy of random groups equating with very small samples. *Journal of Educational Measurement, 42,* 309–330.